Tugas Pekan ke-3 Model Bahasa / Semantik Vektor I

Batas pengumpulan: Jumat 2 Oktober 2020, pukul 10.59 pagi, melalui LMS

Deskripsi

Buatlah matriks Tf-IDF, matriks co-occurrence term-context, dan PPMI dari 20 artikel berbahasa Indonesia, dengan 10 artikel terkait dengan sebuah topik (misal topik A), dan 10 sisanya terkait dengan topik yang berbeda (misal topik B). Usahakan topik A dan topik B cukup berbeda (tergambar pada kosa kata yang ada di dua kelompok artikel tersebut).

Setelah Anda membuat matriks Tf-IDF dan PPMI, lakukan eksperimen sebagai berikut:

- 1. Hitung berapa persen elemen matriks Tf-IDF dan PPMI yang bernilai tidak sama dengan 0.
- 2. **Berdasar matriks TF-IDF**. Hitung nilai *cosine similarity* **antar dokumen** dengan topik yang sama, dan antar dokumen dengan topik yang berbeda. Beri analisis singkat terhadap nilai cosine similarity yang diperoleh, apakah sesuai dengan yang seharusnya?
- 3. **Berdasar matriks TF-IDF**. Hitung nilai *cosine similarity* **antar kata,** dengan contoh pasangan kata yang berasal dari dokumen dengan topik yang sama dan contoh pasangan kata yang berasal dari dokumen dengan topik berbeda. Beri analisis singkat terhadap nilai *cosine similarity* yang diperoleh.
- 4. **Berdasar matriks** *co-occurrence term-context*. Lakukan perhitungan nilai *cosine similarity* **antar kata** seperti poin nomor 3. Bandingkan hasil yang diperoleh dengan hasil dari eksperimen nomor 3.
- 5. **Berdasar matriks PPMI**, periksa nilai PPMI antar kata sesuai dengan yang digunakan pada eksperimen nomor 3 dan 4. Bandingkan hasil yang diperoleh dengan hasil dari eksperimen nomor 3 dan 4.

Informasi yang harus dituliskan pada laporan:

- 1. Keterangan pemilihan topik artikel dan sumber artikel.
- 2. Ukuran matriks Tf-IDF, co-occurrence term-context, dan PPMI.
- 3. Jawaban dari pertanyaan eksperimen yang sudah disebutkan sebelumnya.

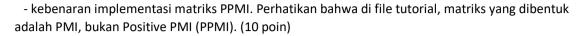
File yang harus dikumpulkan:

- 1. Program dan kelengkapannya: 1 file kode program python (.py) + 20 file artikel (.txt) + petunjuk menjalankan program (.txt).
- 2. Laporan: 1 file pdf, maksimum panjang laporan adalah 2 halaman.

Penilaian: 40% source code + 60% laporan

Detail penilaian:

- a. Program:
- kebenaran implementasi pembacaan file masukan dan pembangunan matriks Tf-IDF dan co-occurrence term-context. (20 poin)
 - kebenaran implementasi cosine similarity antar vektor kata dari matriks Tf-IDF. (10 poin)



- b. Laporan:
 - kelengkapan (30 poin)
 - analisis (30 poin)

Jika ada pertanyaan, silakan disampaikan melalui *channel* pekan_3_tugas_vektor_semantik_1 di slack.